KEYWORD GENERATION FOR LYRICS

Bin Wei
U. Rochester
Comp. Sci. Dept.

Chengliang Zhang
U. Rochester
Comp. Sci. Dept.

Mitsunori Ogihara
U. Rochester
Comp. Sci. Dept.

ABSTRACT

This paper proposes a scheme for content based keyword generation of song lyrics. Syntactic as well semantic similarity is used for sentence level clustering to separate the topic from the background of a song. A method is proposed to search for a center in the semantic graph of WordNet for generating keywords not contained in original text.

1 INTRODUCTION

The growth of Web lyrics databases such as lyrics.com and absolutelyric.com raises the issue of processing and understanding lyrics automatically. However, previous work (see [4]) suggests that lyrics are tough for natural language processing. There are three issues that are specific to keyword extraction, First, the "correct" outcome is usually subtle. Unlike other text data such as news, topic words in lyrics have a very small number of occurrences. Instead, a large portion will be devoted to the background. This makes the frequency information not so useful, or even misleading. Second, because lyrics are free-style, approaches based on word positions also face difficulty. Third, for some lyrics, meaningful keywords may not even be included in the original text. Some sort of induction is needed to find suitable keywords.

To handle the first two issues, we propose to use a sentence-level clustering so as to separate the topic from the background and eliminate the position information irrelevant. To handle the third issues, we propose to uses various relation links of WordNet [1], assuming that the center of the discovered links serves as the central keyword. Keywords generated for each intra-lyric cluster are compared among lyrics, enabling separation of lyrics on similar topics with different backgrounds.

Our use of WordNet is more comprehensive than the existing WordNet lexical-cohesion-based approach of (Silber and McCoy [5]) called Lexical Chain, in that we not only focus on nouns and concept hierarchy but try to combine verbs and adjectives and logical relations in the analysis. The sentence-level clustering enables us to constrain our search in a compact subgraph of WordNet rather than the whole net. Also, our analysis is different in that the goal is not only to find the relations among words in the lyrics but to introduce new words as the suitable keywords that may not necessarily occur in the original text.

[1] http://WordNet.princeton.edu

2 KEYWORD GENERATION

We use the Stanford parser for obtaining dependencies, where a dependency is a triple of the form (relation, governor, dependent). We then use the Lesk measurement for word similarity inthe WordNet, and the algorithm in [2] for word sense disambiguation.

2.1 Sentence Level Clustering

After the previous steps, we obtain a set of senses and a set of dependencies for a given sentence. To combine the similarity of words in the sentence while preserving the semantics, we view the dependency as a proper unit of meaning, as most phases can be represented in this form. We first define the similarity between two dependencies as:

$$Sim(D_a, D_b) = \sum_{W_i \in D_a, W_j \in D_b} Sim(W_i, W_j).$$

We then view a sentence as a collection of meaning units. The similarity between sentences is thus defined as:

$$Sim(S_a, S_b) = \max_{D_i \in S_a, D_j \in S_b} Sim(D_i, D_j).$$

After obtaining pairwise sentence similarity, we use the data-driven clustering algorithm in [1]. The algorithm accepts a normalized distance matrix P as input, and performs a random walk, with transition matrix P, either until convergence or for a specified number of steps. It can automatically decide the number of clusters by optimizing spectral properties of P.

2.2 Candidate Keywords Selection

We then select candidate keywords inside the cluster. Since lyrics are often not well-formed and thus nouns alone do not offer adequate information, we take adjectives, adverbs, and verbs into consideration, too. The ranking of the words inside a cluster is based on their similarity to other members of the cluster:

$$Score(W_i) = \sum_{W_j \in C_k} Sim(W_i, W_j)$$

Based on this score, we keep only top n candidates for the next step (we set $n = 10$ in our work).

2.3 Candidate Generation and Final Selection

WordNet is a freely available electronic dictionary. In WordNet, each node stands for a set of synonyms, and nodes are connected through links that represent semantic relationships. These relations enable us to find the words logically implied by the original text (for example, we may get the cause of a certain outcome by following the cause link), which solves the "hidden keyword" issue. We thus assume that if a sense is reached by many words in the original text, it is a good candidate for keyword.

The algorithm takes as input a set of words from the lyrics, and executes breadth-first traversal on each word for a given number of steps. To cross the boundary of POS, we also view the glossary field as a kind of link. We parse each glossary and extract the words that turn up as subjects, verbs, or objects. We then simply choose their most common senses and think of these senses as connected with the original sense. The words reachable from a significant percentage(30%) of words will be added to the original word list as a candidate.

2.4 Cluster Selection

Our last step is to pick out a cluster for each lyric to stand for the whole. In other words, we need to separate the parts concerning the topic from the rest. We assume that the topics of a lyric are shared with other lyrics and thus commonly appear, while the background cluster is specific to each song and thus is not so common. Thus we use the degree of overlap for selecting a topic cluster. Let c_{ij} denote the j-th cluster for the i-th lyric and w_{ij} its keywods. The score for c_{ij} is:

$$\sigma_{ij} = \frac{\sum_{k \in T} \max_{\ell} \|c_{ij} \cap c_{k\ell}\|}{\|T\|},$$

where T is the set of k such that for some ℓ the set $w_{k\ell}$ has at least three keywords with c_{ij}.

3 EVALUATION

We perform our test on DigiTrad [2], a database that contains approximately 8,000 folk songs and provides content based (like "school", "marriage", etc) keywords. For the first test, we run our program on the same data sets used in [4]: SONG1 consisting of 408 songs composed of two clusters, labeled "murder" and "marriage" respectively in DigiTrad, and SONG2 consisting of 424 songs, labeled "political" and "religion". We first extract the top 5 words using our approach. We then perform a simple unsupervised clustering based on the semantic similarity between sets of keywords, using the same data-driven approach in [1]. We then compare the error rate against the semantic rule learner work in [4]. We observe that our performance offers improvement in accuracy. Given that our method is unsupervised, while the previous method is supervised, we consider this as a substantial improvement.

[2] http://www.mudcat.org/download.cfm

Method	Data	Balance	Error
Scott&Matwin	SONG1	200/224	30.23%
Proposed	SONG1	195/212	28.14%
Scott&Matwin	SONG2	194/238	32.64%
Proposed	SONG2	200/224	31.22%

Table 1. The comparison between two methods. The third column shows the split between the two clusters.

Label	Size	Hit in Top 5	Hit in Any
marriage	195	47	64
murder	212	50	78
poverty	94	25	35
school	29	9	11

Table 2. The result of topic keywords hits.

Another test is performed by directly counting the number of hits our keywords make in the keywords provided in DigiTrad. We generate the top 5 keywords both for the lyric and for each clusters of the lyric, and check whether the given label is contained. For the same lyric, multiple hits from different clusters are counted as one. Considering that the labels given are general and we allow keywords of different POS, we accept the different forms of the same word (such as "marriage" and "marry"), and words with very similar meanings (such as "school" versus "college"/"university"). For some lyrics, we succeed in separating the topics from noise after the sentence level clustering, but we fail in getting the right cluster.

4 FUTURE WORK

The potential sources of error are the word sense disambiguation and words/links missing from WordNet. We may be able to improve the performance by using semi-supervised methods. Also, more sophisticated word sense disambiguation could produce more accurate result. The use of word co-occurrence in discussion forums and reviews may help in find more words.

5 REFERENCES

[1] A. Azran and Z. Ghahramani. In *Proc. 23rd ICML*, pages 57–64, 2006.

[2] S. Patwardhan, S. Banerjee, and T. Pedersen. In *Proc. 4th CICLing*, pages 241–257, 2003.

[3] T. Pedersen, S. Patwardhan, and J. Michelizzi. In *Proc. 19th AAAI*, pages 1024–1025, 2004.

[4] S. Scott and S. Matwin. In *COLING-ACL 98 Workshop*, pages 38–44, 1998.

[5] H. G. Silber and K. F. McCoy. *Comput. Ling.*, 28(4):487–496, 2002.